Formal Mathematical Framework for Lexical Encoding in Chatbot Design

Elsonlian Neihguk

April 22, 2025

Abstract

This paper presents a novel encoding method for mapping natural language words to numerical representations using positional digit decomposition and letter-weight multiplication. The framework serves as a foundation for lightweight, rule-based chatbot systems.

Chapter 1

Definitions

Let:

- $\Sigma = \{A, B, \ldots, Z\}$ be the alphabet.
- $W = \{w_1, w_2, \ldots, w_n\}$ be a corpus of n words.
- $W \to \mathbb{N}$ assign a numerical value to each word, denoted $V(w)$.

Chapter 2

Letter Weight Assignment

2.1 Frequency Calculation

For each letter $\sigma \in \Sigma$, define its weight $\phi(\sigma)$ as:

$$\phi(\sigma) = \sum_{\substack{w \in W \\ \sigma \in w}} V(w)$$

2.2 Rank-Based Normalization

Assign ordinal ranks $r(\sigma_i)$ sorted by $\phi(\sigma)$:

$$r(\sigma_i) = \begin{cases} 26 & \text{if } \phi(\sigma_i) \text{ is maximal} \\ 1 & \text{if } \phi(\sigma_i) \text{ is minimal} \end{cases}$$

Chapter 3

Word Encoding

3.1 Digit Position Decomposition

For a word $w = \sigma_1\sigma_2\ldots\sigma_k$ with $V(w) = N$, express N in base 10 as:

$$N = \sum_{i=1}^{k} d_i \cdot 10^{p_i} \quad \text{where } d_i \in \{0, ..., 9\},\ p_i \in \mathbb{Z}_{\geq 0}$$

3.2 Position-Weighted Encoding

The encoded value $E(w)$ is computed as:

$$E(w) = \sum_{i=1}^{k} r(\sigma_i) \cdot d_i \cdot 10^{p_i}$$

Example for 'HELLO':

$$E(\text{HELLO}) = r(H) \cdot 0 + r(E) \cdot 0 + r(L) \cdot 400 + r(L) \cdot 6000 + r(O) \cdot 80000$$

Chapter 4

Chatbot Response Mapping

Define $R : \mathbb{N} \to \mathcal{R}$, where $\mathcal{R}$ is a set of responses:

$$R(E(w)) = \begin{cases} \text{"Hello"} & \text{if } E(w) = 1{,}276{,}800 \\ \text{"Goodbye"} & \text{if } E(w) = 950{,}000 \\ \vdots & \vdots \\ \text{"Unknown"} & \text{otherwise} \end{cases}$$

Chapter 5

Theoretical Properties

5.1 Injectivity Analysis

The map $E : W \to \mathbb{N}$ is:

- Non-injective in general (collisions possible).
- Injective if for all w_i, w_j, the term $\sum r(\sigma) \cdot d_i \cdot 10^{p_i}$ is unique.

5.2 Computational Complexity

- Encoding a word w of length k: $\mathcal{O}(k)$ time.
- Storage: $\mathcal{O}(n)$ for the response map R.

Chapter 6

Extensions

6.1 Contextual Weighting

Augment ϕ with positional weights:

$$\phi'(\sigma, i) = \alpha^i \cdot \phi(\sigma) \quad \text{where } \alpha > 1 \text{ is a decay factor}$$

6.2 Algebraic Generalization

Replace base-10 decomposition with an arbitrary base B:

$$N = \sum_{i=1}^{k} d_i \cdot B^{p_i}, \quad d_i \in \{0, ..., B-1\}$$

Conclusion

This framework provides a mathematically rigorous encoding scheme for lexical data in resource-constrained chatbot systems. Future work may explore:

- Topological properties of the encoding space.
- Integration with probabilistic language models.